Technical Disclosure Commons

Defensive Publications Series

November 10, 2017

Detecting modifications to images

N/A

Detecting modifications to images

ABSTRACT

The present disclosure describes techniques to detect whether an image, e.g., a photograph, has been modified. For example, image modification can include application of one or more filters to the image. Further, the techniques described quantify the degree of change in the image due to the applied filters. One or more machine-learned models are trained and applied to determine the types of filters that are likely applied to an image. Training data for the models include unmodified images and the same images after one or more filters have been applied. The models are trained to determine the degree to which the image has been modified and to output probabilities that an image has been subjected to particular types of filters.

KEYWORDS

- image filter
- photo effect
- image forensics
- Nofilter
- image analysis
- image authentication

BACKGROUND

Images shared on social media or via messaging applications are often modified, e.g., by application of image filters, prior to the sharing. A viewer of an image may wonder whether a shared image is realistic, and the degree to which it has been modified.

DESCRIPTION

Filters applied to an image are typically predefined within image-manipulation applications. For example, some example filters include sepia (that changes the image to a sepia tinted version), vintage (that mimics the appearance of an old photograph), vignette, etc. Techniques of this disclosure detect the predefined filter(s), or combinations of filters, that have likely been applied to an image. Further, the techniques also provide a confidence score for the detection.

One or more machine-learned models are trained and applied to determine the types of filters that are likely applied to an image. The output can also quantify the degree of change to the image, e.g., "superficial change," "substantial change," "far from reality," etc.

Training data for the models include unmodified images and the same images after one or more filters have been applied. Training data can include unmodified images captured with a variety of devices, and the same images after augmentation or modification with combinations of filters. Filters that are used to generate modified images used during training include commonly available filters, e.g., vintage, vignette, sepia, contrast increase, etc.

The models are also trained to determine the degree to which the image has been modified. The models are also trained to produce as output probabilities that an image has been subjected to filters of particular types. The machine learning models are trained to differentiate between deliberate user-induced image filtering and device-specific image processing that is done by default. The models respond to default device-specific image processing by producing a neutral output. One or more of the models can be implemented, e.g., as a convolutional neural network, followed by a feed forward neural network.

During the training phase, a general score quantifying the degree of modification is determined, e.g., as the mean-squared distance of the original image from the modified image. The model is trained to generate several outputs, such as:

- for supported filters or combinations thereof, a prediction of the degree to which the filter(s) were applied, along with a confidence score for the prediction;
- a general score quantifying the degree of modification of the image;
- an overall confidence of the predictions based on previous prediction scores; etc.

Further, models can be trained to detect other modifications of an image, e.g., wherein objects or persons within the image are perturbed in a realistic manner. More complex machine learning techniques, e.g., adversarial training, can also be used in such cases.

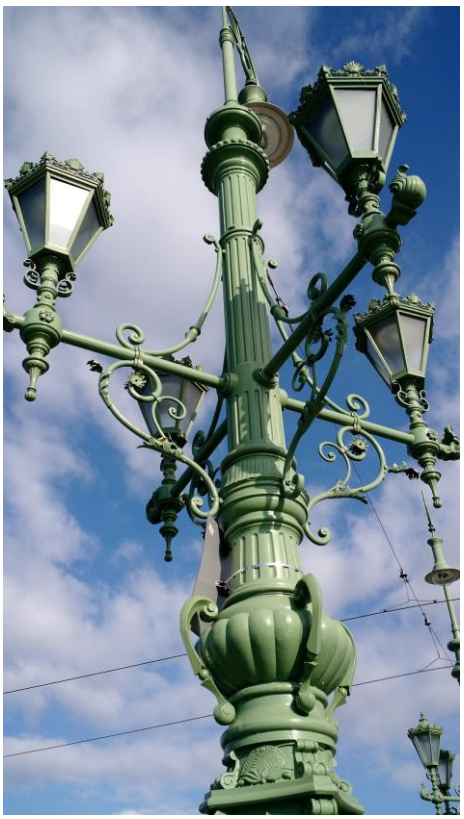
The feature to detect modifications is provided to the user via, e.g., an operating system, within a software application, as screen overlay, etc. For example, the user is enabled to highlight an image and quickly get some filtering statistics about the image, e.g., an output of the form: “3 filters (vintage, vignette and contrast increase); superficial change; 98% confidence.”

Techniques disclosed herein can be used to assess the fidelity of an image, and have applicability in image forensics, social media platforms, messaging applications, photo sharing applications, etc. The techniques can also be used to determine whether an image was taken by a phone or professional camera, to determine filters applied to an image such that a user can produce a similar effect on another image, to determine a filter or image modification application to use to achieve a particular effect, etc.

While the foregoing discussion refers to trained machine-learning models, the detection of modifications applied to the image can also be performed using other techniques, e.g., heuristics that support a limited set of filters. However, such techniques may be complicated and

limited in terms of the modifications that can be detected.

Example



Original image

Modified image

Fig. 1: Example of detection of image modification

Fig. 1 shows an example original image and a corresponding modified image. The techniques described herein analyze the modified image, e.g., using one or more trained machine-learning models, to determine that the image has been modified. For example, the techniques determine that the modified image of Fig. 1 has two filters applied to it, a sepia filter and a vignette.

CONCLUSION

The present disclosure describes techniques to detect whether an image, e.g., a photograph, has been modified. For example, image modification can include application of one or more filters to the image. Further, the techniques described quantify the degree of change in the image due to the applied filters. One or more machine-learned models are trained and applied to determine the types of filters that are likely applied to an image. Training data for the models include unmodified images and the same images after one or more filters have been applied. The models are trained to determine the degree to which the image has been modified and to output probabilities that an image has been subjected to particular types of filters. Techniques disclosed herein can be used to assess the fidelity of an image, and have applicability in various context such as social media platforms, messaging applications, photo sharing applications, virtual assistants, etc.